DOCUMENT RESUME

ED 177 194 TM 009 687

AUTHOR Bernknopf, Stan; And Others
TITLE A Defensible Model for Determining a Minimal Cut-Off Score for Criterion Referenced Tests.
PUB DATE Apr 79
NOTE 28p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Complexity Level; Counselor Certification; Court Litigation; *Criterion Referenced Tests; *Cutting Scores; Guessing (Tests); Higher Education; *Knowledge Level; Mastery Tests; *Mathematical Models; Minimum Competency Testing; Occupational Tests; *Predictive Validity; Test Items; *Test Reliability

ABSTRACT

The effectiveness of a model for determining a minimal cut-off score for criterion-referenced tests was examined. The model, based upon techniques presented originally by Nedelsky and by Angoff, was first used in conjunction with a multiple choice test developed for use in certifying school counselors in Georgia. A "knowledge estimation panel" was used to determine criterion scores; the panel of subject experts estimates the percentages of minimally knowledgeable candidates who could be expected to know (not guess) the correct answers. Nineteen judges on the panel rated 684 items. The resulting criterion scores were then lowered to adjust for standard error and raised to adjust for guessing. The consistency of results was supported by a comparison of parallel test forms and comparable halves of the panel. The panel's estimates of item difficulty were found to be highly correlated with actual difficulty when trial tests were administered. Beta coefficients of agreement also supported the consistency of mastery decisions made. (The legal importance of the models' validity is briefly discussed, and several references to court cases are appended). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177194

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A DEFENSIBLE MODEL FOR DETERMINING A MINIMAL CUT-OFF SCORE FOR CRITERION REFERENCED TESTS

By

Stan Bernknopf
GEORGIA STATE DEPARTMENT OF EDUCATION

Allen Curry
Systems Oriented Measurement and Analysis

W. L. Bashaw
University of Georgia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Stanley Bernknopf
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

TM009 687

Abstract

The effectiveness of a model for determining a minimal cut-off score for criterion-referenced tests is examined in this paper. The model employs a "knowledge estimation panel" and is based on the ability to predict performance for a specific group of examinees. The model is evaluated in terms of consistency of results and in relation to observed score test performance. The validity of the model is discussed from a statistical and legal framework. The latter involves a brief summary of recent court decisions and the influence of these decisions on the model.

A Defensible Model for Determining a Minimal Cut-Off Score for Criterion Referenced Tests

Introduction

Over the course of the past decade educators have witnessed the emergence of the competency-based movement. The movement has grown in leaps and bounds and has affected all aspects of the public and private sectors. Initial efforts of the competency-based movement dealt with student mastery at the classroom level. From there the emphasis shifted to teacher education in the form of competency-based preparation. Competency-based strategies are now being considered by certification agencies and legislators as a means of certifying school personnel. Several states, either through legislative mandate or through state education policy, have adopted both competency-based teacher preparation and performance-based certification procedures.

As more and more states move toward a competency-based certification model, several factors become evident. First, most models now being planned or implemented incorporate the use of criterion-referenced tests. Second, the concept of "Minimal" acceptable performance has replaced the concept of total mastery. Third, a number of recent court decisions relating to validity have direct implications for states that are planning competency-based certification models. These points are by no means mutually exclusive. For example, recent court decisions which have ruled against certifying agencies have done so on the basis of the lack of validity of the instruments and

the choice of cut-off score used in the credentialling process (Georgia Association of Educators, v. J. Nix, 1975; U. S. v. State of North Carolina, 1975, 1977). Specifically, the courts have been concerned about evidence which relates a given cut-off score to a specific level of job performance which can be labeled "minimally acceptable."

Validity is an all encompassing term which relates to many aspects of test development. However, in view of recent court decisions, it is of primary concern when dealing with cut-off score justification. Two points must be emphasized. First, unlike norm referenced tests, criterion-referenced tests used as a part of a credentialling model must yield a score which is interpretable with regard to a persons "functional" level of performance. This implies that the cut-off score is determined independently of a norm group's test performance. Second, it is not the intent of the test user to discriminate between applicants as to degree of proficiency. The intent is to discriminate only between those who have reached a required minimal level of performance and those who have no.. This statement implies that a minimal level of test performance can be defined and related to a defined minimal level of job performance.

The influence of these two points is evident in the following statement which summarizes several court decisions 'While states have the right to set minimal qualification standards, the standard, in order to be considered valid must be job-related and logical (Dent v. West Virginia, 1899; U. S. v. State of North Carolina, 1975; U. S. v. State of South Carolina, 1977). A prime example of the court's interpretation comes form a 1976 Georgia case (Georgia

Association of Educators v. J. Nix, 1975). The case involved the use of the National Teacher Examination (NTE) for certification of teachers. While the court ruled that Georgia has the right to determine the minimum level of competence which shall be required of its teachers, it ruled that the required score of 1225 was chosen without considering what level of academic knowledge would be minimal for competent teaching. A similar decision was reached by the court in a 1975 Mississippi case (Armstead v. Starkville Municipal Separate School District, 1975). In this instance the Starkville Municipal Separate School District required that incumbent teachers and future applicants achieve a minimum score on the Graduate Record Examination (GRE). The court ruled that the cut-off score requirement was not reasonably related to the purpose for which it was to be used.

Procedure

Within this framework, a model for determining a minimal cut-off score for criterion-referenced tests was developed. The model is based on techniques employed by an ESEA Title IV project which was commissioned to develop a criterion-referenced test which would be used as part of the certification requirement for school counselors in Georgia.[1] The model is presented as an example of a simple, straight forward procedure which is valid and defensible. This paper will report an empirical application of the model along with various procedures used to examine its utility.

1. The technique discussed is based on models presented originally by Nedelsky (1954) and Angoff (1971) and a modified Angoff model used by ETS (1976) in a study for South Carolina.

The process used to determine the criterion scores called for the creation of a "Knowledge Estimation Panel." The panel, consisting of content experts, provided estimates (predictions) of the percentages of minimally knowledgeable candidates for initial certification who would be expected to know the answers to the questions on the exam. These predicted percentages can be accumulated over items to yield cutting score estimates needed in order to be considered minimally knowledgeable. Thus, this procedure establishes item content rather than examinee performance as the basis for determining cutting scores. The approach provides one of the best estimates of the probability of classifying examinees into correct mastery or nonmastery status (Reechman and Oosterhoff, 1976).

The practice of using a knowledge estimation panel is clearly an alternative to actual identification of candidates who have been judged on independent grounds, to have minimally acceptable amount of knowledge. If such a group were identified, their observed score distributions obtained by administering the appropriate tests could be used for developing the criterion scores. The problem with such a procedure lies with the identification of the group.

The procedure of using human judgements as the raw data from which to derive the needed information is sometimes questioned. However, the professional literature has demonstrated the usefulness and reliability of such a method in many cases (Bratfisch, Borg and Dornic, 1972; Large and Diamond, 1954; Tinkelman, 1947; Quereshi and Fisher, 1977).

The task of the panel was to estimate the percentage of minimally knowledgeable candidates who would know the answers to individual test questions. Members of the panel were required, first, to develop a concept of a hypothetical reference group and, second, to make specific judgments about the knowledge possessed by the members of that group. The judges were asked, individually, to "draw upon their experience to construct a hypothetical group of persons, each of whom, in their judgment, has the minimum amount of academic knowledge to perform effectively in the schools." The panel members were then instructed to "assume that the candidates would not guess blindly if they did not know the answer and to estimate the percentage of the candidates who would know the answer."

A nine point scale of percentage values (15, 25, 35, 45, 55, 65, 75, 85, 95) was used. Verbal descriptors were not used in order to avoid the possibility of biasing the judgments. The scale was centered about 55, a point that approximates the average difficulty based on previous test data.

Six hundred and eighty-four items were judged by 19 participants. The answer to each question was shown. This prevented the judges from being placed in a test-taking situation rather than a difficulty estimating situation. Also, not all judges could be expected to know the correct answers to all questions and it is reasonable to assume that a judge's estimate can be distorted by the actual difficulty the judge has in answering a question.

Analysis

The first step in the analysis of the judges' estimates was to tablulate for each test item the mean percentage, the standard deviation, and the standard error of the mean. One could sum the average percentage for the items included in a given test and consider this sum as a criterion score. This is possible because the mean of a set of test scores is equal to the sum of the mean scores on the individual questions, and each percentage is an estimate of the mean score on a question ($\Sigma p = \bar{x}$). This procedure is possible under the conditions that: (1) judges were asked to estimate the percentages of candidates who know the answers to the questions, not the percentages who would select the correct answer; and (2) estimates are based on a hypothetically defined group of minimally knowledgeable candidates. While this procedure is sound for obtaining an estimated criterion score, it does not take into account the possibility of making an incorrect decision or the effect of guessing on the part of the examinee.

In a criterion-referenced approach to assessment, we consider an examinee competent (at some defined level) if the candidate is able to answer at least a given percentage of items in a population of items. This minimum percentage reflects the degree of mastery judged sufficient for the objective. Implicit in the procedure is the fact that the mastery, non-mastery decision is based on a small sample of observations (test items) and, as a result, includes some error in the decision process. This error can lead to incorrectly

passing or failing an examinee. It is unlikely that such types of error will occur for examinees who obtain scores far below or far above the criterion score. However, it is likely that such errors will occur for those examinees obtaining scores close to the criterion score. The test developer may choose to manipulate the criterion-score in such a manner as to decrease the probability of making one type of decision error while increasing the probability of making an error in the opposite direction.

The decision as to which way to adjust the criterion score is based on the purpose of the test and the relative seriousness of the two decision errors involved. In the present example, a decision was made to adjust the criterion-score in such a manner as to decrease the probability of not accepting an examinee who is competent while increasing the probability of accepting an examinee who is not competent. The overriding factor contributing to this decision was the fact that the exam is only one of several variables that will be used during the certification process. As such, there is additional opportunity to detect an error made through the criterion-referenced measurement process. Moreover, the choice of eliminating only candidates who score significantly poor is in keeping with standard policy that certifying agencies ought to be concerned primarily with minimal standards.

In order to account for our decision concerning which type of error is the more critical, each estimated item percentage was decreased by two standard errors of the mean percentage.[1] These reduced percentages were summed over the items included in a given

test. This resulted in an adjusted criterion-score (C_A) somewhat lower than the criterion-score based on the sums of the unadjusted mean percentages. This procedure can be expressed by two formulas:

$$P'_i = \bar{x}_i - 2\sigma(\bar{x}_i) \qquad (1)$$

where $\bar{x}_i$ = the mean percentage estimate for item i

$\sigma(\bar{x}_i)$ = the standard error of the mean percentage for item i, and

P'_i = the adjusted percentage estimated of item i

and

$$C_A + \sum_{i=1}^{I} P'_i$$

where C_A is the criterion-score for a given competency adjusted for decision error and the sum is over all items which measure a given competency[2].

The decision regarding what proportion of standard error to subtract was based on examination of the various confidence intervals produced by differing proportions. The decision to adjust the criterion-score at the item level (as opposed to the criterion-score level) was made on the basis of item utility. That is, once P_i is calculated for a given item, it can be used in combination with any other item in future forms of the test.

The next consideration is the effect of guessing. The task of the panel was to estimate the percentages of examinees who <u>know</u> the answers to the questions, <u>not</u> the percentages who would select the correct answer. As previously mentioned, this is a necessary requirement in order to make the above procedure valid. In order to compensate for guessing, a further adjustment had to be made to the criterion-score. Since each item is multiple choice with four

alternatives, it is reasonable to assume that 25 percent of the items not known by minimally competent examinees would be answered correctly by chance. Therefore, the correction for guessing that was chosen is to raise the adjusted criterion-score (C_A) by 25 percent of the remaining items. This yields a new criterion-score which is also adjusted for guessing (C_B). This relationship can be expressed as

$$C_B = C_A + .25\ (I - C_A) \qquad (3)$$

where

C_A = the criterion score for a given competency adjusted for decision error,

I = the number of items measuring the competency, and

C_B = the criterion score for a given competency adjusted for decision error and guessing.

Results

The consistency of the results obtained from the process was analyzed by comparing data from two parallel forms of the test and by comparing data from "half panels" by test form. "Half panels" refers to a division of the total group of judges into two groups. The division is based on specific variables (sex, race, job description, etc.) which insured that the two groups were comparable. This is similar to obtaining data on two separate occasions or from two separate studies.

The comparability of the results for the two parallel forms was examined in terms of the degree of similarity between the judgments. For each form of the test, the unadjusted means, the

standard errors of the means, and the adjusted means were computed by competency. This data was based on full panel estimates. In addition, a final cut-off score, rounded off to the nearest whole item for each form of the test, by competency, was determined. The differences between the standard error, the differences between adjusted means (C_B), the rounded cut-off scores, and the differences between cut-off scores by competency are shown in Table 1.

Insert Table 1 about here

The differences between the two forms for the standard errors and the adjusted means vary from .002 to .047 and from .007 to .712, respectively. These differences are quite small and show a large degree of consistency between forms. In the case of the adjusted means, no difference reached a value equal to one whole item, and 17 of the 20 differences were less than one-half of an item. In terms of actual cut-off scores (estimated adjusted means rounded to the nearest whole number), 12 of the 20 competency scores produced the same value. Since the two forms are parallel in terms of content, and since the basis for determining these values are based on content rather than examinee performance, the data would seem to indicate that the panel's judgments are very consistent and item oriented. A lack of consideration as to the relative difficulty of each item would have produced results which would seem random and yield large differences between forms.

Table 2 shows the results of examining the same data by half panels and by form. For Form A, ten of the 20 scores differ by one item. The remaining ten scores round to the same number of items. Form B shows similar results with eight of the 20 scores differing by one item. The differences between the standard errors vary from .002 to .754 and from .012 to .781 for Forms A and B, respectively. The corresponding differences between the adjusted mean scores vary from .004 to .957 for Form A and from .010 to 1.059 for Form B.

Insert Table 2 about here

The size of the difference between half panels for the various statistics presented reflects the degree of consistency displayed at the competency level. It represents the difference in the results that would be obtained in a study based only on Panel One and a study based only on Panel Two. In such a case it is possible that during the calculations of the differences between half panel means, a positive difference for one item could have been cancelled by a negative difference of the same size for another item. In order to examine the degree of consistency at the item level, where positive and negative differences of the same size were not allowed to cancel each other, correlations were calculated between half panels by competency. That is, for each competency a correlation was computed based on the item means within the competency. These correlations are shown in Table 3 by test form. In addition, the mean correlation from the two forms combined is also presented.

Insert Table 3 about here

The correlations by form show a great amount of consistency between half panels at the item level. In most cases the correlations are quite high and far beyond traditional significance levels. Comparisons of correlation coefficients between test forms should be done with care. In many cases, the closeness of these coefficients is an artifact of the item overlap between forms. Competencies 16 and 18 are the only competencies which contain a completely different set of items in each form. The remaining competencies range from 8 to 78 percent overlap.

Table 4 shows the relative relationship between panel mean estimates and actual observed score mean estimates. The observed score means are based on data obtained from subsequent test administrations. This relationship is expressed as a percentage of total items. It should be noted that, for Form A, the panel estimates which represent a hypothesized minimally competent group exceeded the observed estimates in four competency areas. The corresponding figure for Form B is two areas. Since the observed scores are based on a random sample of examinees as opposed to a minimally competent group, one would expect the obtained values to be higher than the panel estimates.

Insert Table 4 about here

By correlating the two estimates, one can see that the relative estimated difficulty of each competency area corresponds to the actual difficulty. In other words, those areas which the panel saw as most difficult did, in fact, tend to be most difficult. The actual correlations are .74 and .85, for Forms A and B, respectively.

Table 5 provides performance statistics based on the tryout administration of Forms A and B. The table shows observed score means and the percentage of examinees earning scores higher than the mean estimated from the panel judgments (minimal competency score). The value of the score means cannot be compared across competencies since the scores are not on comparable scales. However, the percentage of examinees whose scores exceeded the estimated means can be compared. These values range from 41.5 percent to 95.9 percent for Form A and from 52.9 percent to 90.7 percent for Form B.

Insert Table 5 about here

It should also be noted that two of the observed score means were lower than their corresponding cut-off scores for Form A. For Form B, three observed score means were lower than their corresponding cut-off scores. These cases represent the only areas where mean test performance was below the point corresponding to minimal competency. Within the same framework, it should also be pointed out that for Form A, only one competency had fewer than half the examinees earning scores higher than that judged to characterize the student with a minimum amount of academic knowledge. All competencies in Form B

and the remaining competencies in Form A had more than half the examinees earning scores higher than that judged to characterize the student with a minimum amount of academic knowledge.

As mentioned earlier, the approach employed is designed to provide a basis for correctly classifying an examinee into a mastery or nonmastery status. In order to assess the consistency of the pass/fail decisions resulting from the use of these cut-off scores, a method proposed by Marshall and Haertel (1975) was employed. The method is simply a measure of decision-making consistency which is based on a coefficient of agreement called beta. Their approach considers the amount of agreement which occurs when an instrument is split into two parts, each consisting of the same number of items. Since there are many possible ways in which a test can be split, the procedure involves computing a coefficient of agreement for every possible split of the test. Beta equals the average of all of these coefficients of agreement.[3]

Table 6 presents for each of the 20 competency areas the mean split-half coefficients of agreement (coefficient beta). Since coefficient beta is essentially a half-test index, computed values were stepped up by means of the Spearman-Brown formula so as to approximate the value of the coefficient if it were based on the number of items comprising the full length test. These stepped-up values are presented in parentheses.

Insert Table 6 about here

Conclusions

State education agencies in recent years have begun exerting considerable leadership in the field of teacher education and certification. New and revised procedures for certifying teachers are being planned and instituted. This effort, however, is being threatened by recent court rulings involving questions as to what are or are not legally permissible procedures a state may use as part of its certification process. State education agencies must be aware of these legal rulings, as well as options which are consistent with these rulings.

The technique presented in this paper seems to offer one such option. The essential value of the technique is that it produces a cut-off score that is based on a minimum proficiency standard which is directly related to the competencies that are being measured. Needless to say, the technique assumes a test which has proven validity, especially in terms of its job-relatedness. In such a case, where the test content has been shown to be related to the job in question, the procedure of determining a cut-off score based on item content rather than examinee performance offers both a "rational" and "logical" alternative.

In addition, the procedure is simple to carry out, yields consistent valid results which correspond to expected observed score performance, and offers the opportunity to set the cut-off score in relation to any given level of performance. This last point has great significance for those agencies who plan to use a single test

to issue several levels of certification. While the present study was designed to yield a single cut-off score representing a given level of proficiency, the procedure can easily be modified to produce multiple passing scores. In such a case, each score would be associated with various levels of personnel experience and/or performance.

The data we have presented is from a school counselor certification project. The agreement of half-panels and the agreement of judgments with real data may differ substantially in other areas of application. These agreements will depend upon the accuracy and clarity with which job descriptions and item domains can be specified.

REFERENCE NOTES

1. The choice of "2" corresponds approximately to a normal curve one-tailed test at the .022 level of significance.

2. Equation 2 is possible because $\Sigma P = \bar{x}$.

3. Coefficient beta is conceptually analogous to the more familiar KR-20 internal consistency estimate of reliability. Whereas KR-20 has been shown to represent the mean of all possible split half reliability coefficients, coefficient beta is equal to the mean of all possible split half coefficients of agreement.

REFERENCES

Angoff, W. H., Scales norms, and equivalent scores. In R. L. Thorndike (ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Armstead v. Starknille Municipal Separate School District, 325 F. Supp. 560 (N.D. Miss 1971), Modified, 461 F. 2d. 276 (1972).

Bratfisch, O., Barg, G. and Dornic, G. Perceived Item-Difficulty in Three Tests of Intellectual Performance Capacity (Report No. 29). Reports from the Institute of Applied Psychology, the University of Stockhold, 1972.

Dent v. West Virginia, 129 U.S. 114 (1889).

Educational Testing Service, Report on a study of the use of the National Teachers Examination by the State of South Carolina. Princeton, NJ: Educational Testing Service, 1976.

Georgia Association of Educators v. Nix, 407 F. Supp. 1102 (1976).

Large, I. and Diamond, L. The Prediction of Absolute Item Difficulty by Ranking and Estimating Techniques. Educational and Psychological Measurement, 365-372, 1954.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement. 1954, 14, 3-19.

Quereshi, M.Y. and Fisher, T. L. Logical versus Empirical Estimates of Item Difficulty. Educational and Psychological Measurement, 91-100, 1977.

Reechman, S. L. and Oosterhoff, A. C. Strategy Guidelines for the Construction of Mastery Tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

Tinkelman, S. Difficulty Prediction of Test Items (Teachers College Contributions to Education No. 941) Bureau of Publications, Teachers College, Columbia University, 1947.

United States v. State of North Carolina, 400 F. Supp. 343 (E. D. N. C. 1975), 425 F. Supp. 789 (E. D. N. C. 1977).

United States v. State of South Carolina, 15 FEP Cases 1196 (D. C. S. C. 1977).

TABLE 1

Difference Between Estimated Standard Errors and Estimated Means (C_B)(Form A vs Form B, Full Panel Data)

Competency	K	Absolute Difference of S.E. Form A vs Form B	Absolute Difference of Estimated Means (C_B) Form A vs Form B	Cut-Off Scores Rounded to Nearest Whole Item Form A	Form B	Absolute Difference of Cut-Off Score Form A vs Form B
1	26	.019	.712	16	15	1
2	43	.004	.012	27	27	0
3	23	.018	.188	15	15	0
4	17	.021	.311	11	11	0
5	10	.014	.216	7	7	0
6	54	.015	.465	34	35	1
7	12	.002	.132	6	7	1
8	14	.014	.007	9	9	0
9	8	.004	.058	5	5	0
10	16	.047	.221	10	11	1
11	10	.005	.144	7	7	0
12	12	.020	.306	8	7	1
13	12	.014	.019	8	8	0
14	10	.019	.085	7	7	0
15	10	.011	.180	7	6	1
16	13	.017	.045	8	8	0
17	16	.021	.668	10	11	1
18	9	.019	.058	6	6	0
19	9	.005	.128	6	6	0
20	20	.009	.591	13	12	1

TABLE 2

Difference Between Estimated Standard Errors and Estimated Means (C_B) by Test Forms

(Half Panel 1 vs Half Panel 2)

	Form A			Form B		
Competency	Size of Difference Between Estimated Standard Error	Size of Difference Between Estimated Means (C_B)	Absolute Difference of Cut-Off Score	Size of Difference Between Estimated Standard Error	Size of Difference Between Estimated Means (C_B)	Absolute Difference of Cut-Off Score
1	.496	.469	1	.506	.469	0
2	.553	.687	1	.765	1.059	1
3	.235	.056	0	.231	.022	0
4	.171	.004	0	.271	.096	0
5	.155	.149	0	.014	.160	0
6	.754	.957	1	.781	.347	0
7	.151	.570	1	.166	.409	1
8	.155	.915	1	.105	.836	1
9	.101	.152	0	.110	.188	0
10	.034	.142	0	.128	.010	0
11	.047	.163	0	.017	.077	0
12	.090	.502	1	.097	.781	0
13	.047	.300	0	.152	.098	0
14	.070	.432	1	.012	.398	1
15	.148	.832	1	.102	.734	1
16	.207	.419	1	.208	.728	1
17	.243	.494	1	.165	.671	1
18	.074	.275	0	.051	.210	0
19	.002	.089	0	.056	.278	0
20	.166	.692	0	.165	.401	1

TABLE 3

Correlations Between Half-Panels by
Competency and Test Form

Competency	K	% Common Items	Form A		Form B		Mean	
			r	r^2	r	r^2	r	r^2
1	26	65.4	.83***	.69	.81***	.66	.82***	.67
2	43	58.1	.85***	.72	.86***	.73	.86***	.73
3	23	62.5	.72***	.52	.81***	.65	.77***	.59
4	17	11.8	.84***	.70	.86***	.73	.85***	.72
5	10	10.0	.54	.29	.79**	.62	.69**	.48
6	54	58.5	.66***	.44	.69***	.47	.68***	.46
7	12	41.6	.52*	.27	.70**	.48	.62*	.38
8	14	78.6	.83***	.70	.82***	.68	.83***	.70
9	8	75.0	.35	.12	.57	.33	.47	.22
10	16	37.5	.51*	.26	.79***	.62	.68***	.46
11	10	40.0	.69*	.48	.77**	.60	.73***	.53
12	12	16.7	.78**	.61	.85***	.73	.83***	.67
13	12	8.3	.52*	.27	.37	.13	.45	.20
14	10	50.0	.83**	.69	.62*	.38	.74*	.55
15	10	50.0	.71*	.50	.77**	.60	.74*	.55
16	13	30.8	.77**	.59	.73**	.53	.75**	.56
17	16	0.0	.82***	.68	.72***	.51	.78***	.61
18	9	0.0	.91***	.82	.76**	.58	.85**	.72
19	9	55.6	.76**	.58	.85**	.72	.81**	.66
20	20	15.0	.91***	.83	.78***	.61	.86***	.74
		43.0	.75***	.56	.77***	.59	.76***	.58

*P<.05
**P<.01
***P<.001

TABLE 4

Relationship Between Panel Mean Estimates and Observed Score Mean Estimates as a Percentage of Total Items by Competency and Test Form

		Percent of Total Items					
		Form A			Form B		
Competency	Number of Items	Panel Estimate	Observed Estimate	Absolute Difference	Panel Estimate	Observed Estimate	Absolute Difference
1	26	62.31	69.27	6.96	59.57	65.54	5.97
2	43	62.07	72.93	10.86	62.09	71.49	9.40
3	23	64.01	71.26	7.25	63.20	73.00	9.80
4	17	63.66	62.94	0.72	61.83	64.59	2.76
5	10	65.63	59.60	6.03	67.79	78.60	10.81
6	54	63.17	67.22	4.05	64.03	69.54	5.51
7	12	53.57	52.67	0.90	54.67	48.33	6.34
8	14	61.08	66.79	5.71	61.03	70.00	8.97
9	8	63.05	63.00	0.05	62.33	61.50	.83
10	16	65.04	76.56	11.52	66.41	78.06	11.65
11	10	68.04	76.10	8.06	69.48	79.60	10.12
12	12	63.79	74.58	10.79	61.24	70.00	8.76
13	12	66.56	67.00	0.44	66.40	75.92	9.52
14	10	66.46	74.20	7.74	67.31	71.70	4.39
15	10	65.75	78.20	12.45	63.95	75.50	11.55
16	13	60.48	62.39	1.91	60.82	68.23	7.41
17	16	62.28	75.94	13.66	66.46	74.19	7.73
18	9	72.20	84.67	12.47	71.56	81.11	9.55
19	9	68.47	77.00	8.53	69.89	74.00	4.11
20	20	62.63	66.80	4.17	59.68	67.60	7.92

Panel estimates are adjusted for error and guessing.

NOTE: Correlation between Panel Estimate and Observed Estimate, Form A, is .74 ($p<.001$)
Correlation between Panel Estimate and Observed Estimate, Form B, is .85 ($p<.001$)

TABLE 5

Percentages of Examinees Above Minimal Competence by Form and Competency

Competency	K	Minimal Competency Score	Form A (N=171)			Form B (N=172)		
				Minimal Competency			Minimal Competency	
			Mean	N	%	Mean	N	%
1	26	16	18.01	141	82.5	17.04	121	70.4
2	43	27	31.36	140	81.9	30.74	136	79.1
3	23	15	16.39	135	79.0	16.79	130	75.6
4	17	11	10.70	95	55.6	10.98	91	52.9
5	10	7	5.96	71	41.5	7.86	135	78.5
6	54	34	36.30	120	70.2	37.55	134	77.9
7	12	6	6.32	115	67.3	5.80	94	54.7
8	14	9	9.35	115	67.3	9.80	125	72.7
9	8	5	5.04	114	66.7	4.92	108	62.8
10	16	11	12.25	138	80.7	12.49	149	86.6
11	10	7	7.61	139	81.3	7.96	148	86.1
12	12	8	8.95	139	81.3	8.40	134	77.9
13	12	8	8.04	107	62.6	9.11	146	84.9
14	10	7	7.42	126	73.7	7.17	120	69.8
15	10	6	7.82	159	93.0	7.55	156	90.7
16	13	8	8.11	115	67.3	8.87	130	75.6
17	16	10	12.15	151	88.3	11.87	148	86.1
18	9	6	7.62	164	95.9	7.30	154	89.5
19	9	6	6.93	143	83.6	6.66	131	76.2
20	20	12	13.36	130	76.0	13.52	135	78.5

TABLE 6

Mean Split-Half Coefficient of Agreement
(Coefficient Beta) By Competency Area

Competency	Number of Items (K)	Minimal Score on Half Test	Coefficient Beta[a] Form A	Form B
1	26	8	.35(.52)	.40(.57)
2	43	14	.46(.64)	.47(.64)
3	23	8	.47(.64)	.60(.76)
4	17	6	.63(.77)	.64(.78)
5	10	4	.61(.76)	.76(.87)
6	54	17	.75(.86)	.81(.90)
7	12	3	.43(.60)	.41(.58)
8	14	5	.50(.66)	.56(.72)
9	8	3	.51(.68)	.52(.68)
10	16	6	.60(.75)	.58(.73)
11	10	4	.54(.71)	.67(.80)
12	12	4	.65(.79)	.56(.72)
13	12	4	.55(.71)	.68(.81)
14	10	4	.57(.72)	.49(.66)
15	10	3	.75(.86)	.76(.86)
16	13	4	.53(.69)	.64(.78)
17	16	5	.66(.80)	.65(.79)
18	9	3	.71(.83)	.65(.79)
19	9	3	.65(.79)	.72(.84)
20	20	6	.46(.64)	.54(.70)

a. Figure in parentheses represents the value of coefficient beta stepped up to K items.